Spring 4-2018

Bootstrapping Forward Selection with Cp

Drew Imhoff
dcimhoff@siu.edu

BOOTSTRAPPING FORWARD SELECTION WITH C_p

by

Drew C. Imhoff

B.S., Southern Illinois University, 2016

A Research Paper
Submitted in Partial Fulfillment of the Requirements for the
Master of Science

Department of Mathematics
in the Graduate School
Southern Illinois University Carbondale
May 2018

RESEARCH PAPER APPROVAL

BOOTSTRAPPING FORWARD SELECTION WITH C_p

by

Drew C. Imhoff

A Research Paper Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Mathematics

Approved by:

David J. Olive

Bhaskar Bhattacharya

Michael Sullivan

Graduate School
Southern Illinois University Carbondale
April 10, 2018

AN ABSTRACT OF THE RESEARCH PAPER OF

DREW C. IMHOFF, for the Master of Science degree in MATHEMATICS, presented APRIL 10, 2018, at Southern Illinois University Carbondale.

TITLE: BOOTSTRAPPING FORWARD SELECTION WITH C_p

MAJOR PROFESSOR: Dr. David J. Olive

This paper presents a method for bootstrapping the multiple linear regression model $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$ using forward selection with the C_p criterion.

KEY WORDS: Bootstrap; Confidence Region; Forward Selection; Prediction Interval.

ACKNOWLEDGMENTS

I would like to take this opportunity to thank my research advisor, Dr. David Olive for overseeing my Master's project and Dr. Bhaskar Bhattacharya and Dr. Michael Sullivan for sitting on my committee. I would also like to thank the professors of Southern Illinois University for their instruction and care over the past two years. A debt of gratitude is owed in particular to Dr. Mary Wright for her guidance and support. Words of thanks are also due to that small circle of fellow students without whose aid my efforts would have faltered long ago. Finally, I would be remiss not to extend my appreciation to my friends and family; their support, encouragement, and strength have brought what seemed beyond reach to within my grasp. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Suppose that the response variable Y_i and at least one predictor variable $x_{i,j}$ are quantitative with $x_{i,1} \equiv 1$. Let $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p})$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ where β_1 corresponds to the intercept. Then the multiple linear regression (MLR) model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1.1}$$

for $i = 1, ..., n$. This model is also called the full model. Here n is the sample size, and assume that the random variables e_i are independent and identically distributed (iid) with variance $V(e_i) = \sigma^2$. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. The ith fitted value $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ and the ith residual $r_i = Y_i - \hat{Y}_i$ where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$. Ordinary least squares (OLS) is often used for inference if n/p is large.

Variable selection is the search for a subset of predictor variables that can be deleted without important loss of information. Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{1.3}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (1.3) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I \tag{1.4}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$.

Forward selection forms a sequence of submodels $I_1, ..., I_M$ where I_j uses j predictors including the constant. Let I_1 use $x_1^* = x_1 \equiv 1$: the model has a constant but no nontrivial predictors. To form I_2, consider all models I with two predictors including x_1^*. Compute $Q_2(I) = SSE(I) = RSS(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let I_2 minimize $Q_2(I)$ for the $p-1$ models I that contain x_1^* and one other predictor. Denote the predictors in I_2 by x_1^*, x_2^*. In general, to form I_j consider all models I with j predictors including variables $x_1^*, ..., x_{j-1}^*$. Compute $Q_j(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let I_j minimize $Q_j(I)$ for the $p - j + 1$ models I that contain $x_1^*, ..., x_{j-1}^*$ and one other predictor not already selected. Denote the predictors in I_j by $x_1^*, ..., x_j^*$. Continue in this manner for $j = 2, ..., p$ where $n \geq 10p$ and p is fixed.

When there is a sequence of p submodels, the final submodel I_d needs to be selected. Let the candidate model I contains a terms, including a constant. For example, let $\boldsymbol{x}_I$ and $\hat{\boldsymbol{\beta}}_I$ be $a \times 1$ vectors. Then there are many criteria used to select the final submodel I_d. For a given data set, p, n, and $\hat{\sigma}^2$ act as constants, and a criterion below may add a constant or be divided by a positive constant without changing the subset I_{min} that minimizes the criterion.

Let criteria $C_S(I)$ have the form

$$C_S(I) = SSE(I) + aK_n\hat{\sigma}^2.$$

These criteria need a good estimator of σ^2. The criterion $C_p(I) = AIC_S(I)$ uses $K_n = 2$ while the $BIC_S(I)$ criterion uses $K_n = \log(n)$. Typically $\hat{\sigma}^2$ is the OLS full model

$$MSE = \sum_{i=1}^n \frac{r_i^2}{n-p}$$

when n/p is large. Then $\hat{\sigma}^2 = MSE$ is a $\sqrt{n}$ consistent estimator of σ^2 under mild conditions by Su and Cook (2012).

The following criterion are described in Burnham and Anderson (2004), but still need n/p large. AIC is due to Akaike (1973) and BIC to Schwarz (1978).

$$AIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + 2a, \quad \text{and}$$

$$BIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + a \log(n).$$

Let I_{min} be the submodel that minimizes the criterion using variable selection with OLS. Following Nishi (1984), the probability that model I_{min} from C_p or AIC underfits goes to zero as $n \to \infty$. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. Since fewer than 2^p regression models I contain the true model, and each such model gives a $\sqrt{n}$ consistent estimator $\hat{\boldsymbol{\beta}}_{I,0}$ of $\boldsymbol{\beta}$, the probability that I_{min} picks one of these models goes to one as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under model (1.3). See Pelawa Watagoda and Olive (2018) and Olive (2017a: p. 123, 2017b: p. 176).

Section 2 considers mixture distributions. Section 3 shows that a bootstrap confidence region can be formed by applying a prediction region to the bootstrap sample, and Section 4 gives a simulation.

CHAPTER 2

MIXTURE DISTRIBUTIONS

Mixture distributions are useful for variable selection since asymptotically $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{I_j,0}$ where $S \subseteq I_j$. See Equation (1.3). A random vector $\boldsymbol{u}$ has a mixture distribution if $\boldsymbol{u}$ equals a random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$.

Definition 1. The distribution of a $g \times 1$ random vector $\boldsymbol{u}$ is a mixture distribution if the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \tag{2.1}$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of a $g \times 1$ random vector $\boldsymbol{u}_j$. Then $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j.

Theorem 1. Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E(h(\boldsymbol{u})) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]. \tag{2.2}$$

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j], \tag{2.3}$$

and $Cov(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$
$\sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$

$$\sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T. \tag{2.4}$$

If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and

$$Cov(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j).$$

This theorem is easy to prove if the $\boldsymbol{u}_j$ are continuous random vectors with (joint) probability density functions (pdfs) $f_{\boldsymbol{u}_j}(\boldsymbol{t})$. Then $\boldsymbol{u}$ is a continuous random vector with pdf

$$f_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j f_{\boldsymbol{u}_j}(\boldsymbol{t}), \quad \text{and}$$

$$E(h(\boldsymbol{u})) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{u}}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^{J} \pi_j \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{u}_j}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]$$

where $E[h(\boldsymbol{u}_j)]$ is the expectation with respect to the random vector $\boldsymbol{u}_j$. Note that

$$E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T. \tag{2.5}$$

Definition 2. The *population mean* of a random $p \times 1$ vector $\boldsymbol{X} = (X_1, ..., X_p)^T$ is

$$E(\boldsymbol{X}) = (E(X_1), ..., E(X_p))^T$$

and the $p \times p$ *population covariance matrix*

$$Cov(\boldsymbol{X}) = E(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T = (\sigma_{ij}).$$

That is, the ij entry of $Cov(\boldsymbol{X})$ is $Cov(X_i, X_j) = \sigma_{ij}$.

Note that $Cov(\boldsymbol{X})$ is a symmetric positive semidefinite matrix. The following results are useful. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \;\text{ and }\; \mathrm{E}(\boldsymbol{X} + \boldsymbol{Y}) = \mathrm{E}(\boldsymbol{X}) + \mathrm{E}(\boldsymbol{Y}) \tag{2.6}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \;\text{ and }\; \mathrm{E}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}\mathrm{E}(\boldsymbol{X})\boldsymbol{B}. \tag{2.7}$$

Thus

$$Cov(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = Cov(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}Cov(\boldsymbol{X})\boldsymbol{A}^T. \tag{2.8}$$

For the multivariate normal (MVN) distribution $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

$$Cov(\boldsymbol{X}) = \boldsymbol{\Sigma}.$$

CHAPTER 3

BOOTSTRAPPING CONFIDENCE REGIONS

Inference will consider bootstrap confidence intervals and bootstrap confidence regions for bootstrap hypothesis testing. Applying the shorth prediction interval and the Olive (2013) prediction region to the bootstrap sample will give the bootstrap confidence intervals and regions.

Consider predicting a future test random variable Z_f given iid training data $Z_1, ..., Z_n$. A large sample $100(1-\delta)\%$ prediction interval (PI) for Z_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Z_f \leq \hat{U}_n) \to 1-\delta$ as the sample size $n \to \infty$. The shorth(c) estimator is useful for making prediction intervals. Let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics of $Z_1, ..., Z_n$. Then let the shortest closed interval containing at least c of the Z_i be

$$\mathrm{shorth(c)} = [\mathrm{Z_{(s)}}, \mathrm{Z_{(s+c-1)}}]. \tag{3.1}$$

Let $\lceil x \rceil$ be the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{3.2}$$

Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) PI has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth($c$) estimator as the large sample $100(1-\delta)\%$ PI where

$$c = \min(n, \lceil n[1-\delta+1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{3.3}$$

Example 1. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76


order data: 76 78 89 111 778
```

```
13 = 89 - 76

        33 = 111 - 78

              689 = 778 - 89
shorth(3) = [76,89]
```

We also want to use bootstrap tests. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Given training data $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$, a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$. For model (1.1), let $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$.

To bootstrap a confidence region, Mahalanobis distances and prediction regions will be useful. Consider predicting a future test value $\boldsymbol{z}_f$, given past training data $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ where the $\boldsymbol{z}_i$ are $g \times 1$ random vectors. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{z}_f \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$. Let the $g \times 1$ column vector T be a multivariate location estimator, and let the $g \times g$ symmetric positive definite matrix $\boldsymbol{C}$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = D^2_{\boldsymbol{z}_i}(T, \boldsymbol{C}) = (\boldsymbol{z}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z}_i - T) \tag{3.4}$$

for each observation $\boldsymbol{z}_i$. Notice that the Euclidean distance of $\boldsymbol{z}_i$ from the estimate of center T is $D_i(T, \boldsymbol{I}_g)$ where $\boldsymbol{I}_g$ is the $g \times g$ identity matrix. The classical Mahalanobis distance D_i uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{z}}, \boldsymbol{S})$, the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{z}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{z}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\mathrm{n}-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{z}_{\mathrm{i}} - \overline{\boldsymbol{z}})(\boldsymbol{z}_{\mathrm{i}} - \overline{\boldsymbol{z}})^{\mathrm{T}}. \tag{3.5}$$

Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + g/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta g/n), \quad \text{otherwise.} \tag{3.6}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let

$$c = \lceil n q_n \rceil. \tag{3.7}$$

Let $(T, \boldsymbol{C}) = (\overline{\boldsymbol{z}}, \boldsymbol{S})$, and let $D_{(U_n)}$ be the $100q_n$th sample quantile of the D_i. Then the Olive (2013) large sample $100(1-\delta)\%$ nonparametric prediction region for a future value $\boldsymbol{z}_f$ given iid data $\boldsymbol{z}_1, ..., , \boldsymbol{z}_n$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{z}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}, \tag{3.8}$$

while the classical large sample $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{z}}, \boldsymbol{S}) \leq \chi^2_{g,1-\delta}\}. \tag{3.9}$$

Definition 3. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 2. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 3. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $Cov(\boldsymbol{w}) = \dfrac{n-1}{n}\boldsymbol{S} \equiv \dfrac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population covariance matrix

$$Cov(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^n \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$Cov(\boldsymbol{w}) = \sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n}\boldsymbol{S}. \ \square$$

Example 3. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(W_i \le y)$$

where the indicator $I(W_i \le y) = 1$ if $W_i \le y$ and $I(W_i \le y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1-F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\theta} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The *empirical bootstrap* or *nonparametric bootstrap* or *naive bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T_{jn}^* is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 4. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)

b1

[1] 3 2 3 2 5 2 6

median(b1)
```

```
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

The bootstrap has been widely used to estimate the population covariance matrix of the statistic $Cov(T_n)$, for testing hypotheses, and for obtaining confidence regions (often confidence intervals). An iid sample $T_{1n}, ..., T_{Bn}$ of size B of the statistic would be very useful for inference, but typically we only have one sample of data and one value $T_n = T_{1n}$ of the statistic. Often $T_n = t(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, and the bootstrap sample $T^*_{1n}, ..., T^*_{Bn}$ is formed where $T^*_{jn} = t(\boldsymbol{w}^*_{j1}, ..., \boldsymbol{w}^*_{jn})$.

The *residual bootstrap* is often useful for additive error regression models of the form $Y_i = m(\boldsymbol{x}_i) + e_i = \hat{m}(\boldsymbol{x}_i) + r_i = \hat{Y}_i + r_i$ for $i = 1, ..., n$ where the ith residual $r_i = Y_i - \hat{Y}_i$. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{r} = (r_1, ..., r_n)^T$, and let $\boldsymbol{X}$ be an $n \times p$ matrix with ith row $\boldsymbol{x}_i^T$. Then the fitted values $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$, and the residuals are obtained by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$. Here the errors e_i are iid, and it would be useful to be able to generate B iid samples $e_{1j}, ..., e_{nj}$ from the distribution of e_i where $j = 1, ..., B$. If the $m(\boldsymbol{x}_i)$ were known, then we could form a vector $\boldsymbol{Y}_j$ where the ith element $Y_{ij} = m(\boldsymbol{x}_i) + e_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}_j$ on $\boldsymbol{X}$. Instead, draw samples $r^*_{1j}, ..., r^*_{nj}$ with replacement from the residuals, then form a vector $\boldsymbol{Y}^*_j$ where the ith element $Y^*_{ij} = \hat{m}(\boldsymbol{x}_i) + r^*_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}^*_j$ on $\boldsymbol{X}$.

The Olive (2017ab, 2018ab) prediction region method obtains a confidence region for $\boldsymbol{\theta}$ by applying the nonparametric prediction region (3.8) to the bootstrap sample $T^*_1, ..., T^*_B$, and the theory for the method is sketched below. Let $\overline{T}^*$ and $\boldsymbol{S}^*_T$ be the sample mean and sample covariance matrix of the bootstrap sample. Assume $n\boldsymbol{S}^*_T \xrightarrow{P} \boldsymbol{\Sigma}_A$. See Machado and Parente (2005) for regularity conditions for this assumption.

Following Bickel and Ren (2001), let the vector of parameters $\boldsymbol{\theta} = T(F)$, the statistic

$T_n = T(F_n)$, and $T^* = T(F_n^*)$ where F is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F_n^* is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (has a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $\boldsymbol{u} = \dot{T}(F)\boldsymbol{z}_F$. Olive (2017b) used these results to show that if $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, then $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \boldsymbol{0}$, $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and that the prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \le D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\} \tag{3.10}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

The prediction region method for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \ne \boldsymbol{\theta}_0$ is simple. Let $\hat{\boldsymbol{\theta}}$ be a consistent estimator of $\boldsymbol{\theta}$ and make a bootstrap sample $\boldsymbol{w}_i = \hat{\boldsymbol{\theta}}_i^* - \boldsymbol{\theta}_0$ for $i = 1, ..., B$. Make the nonparametric prediction region (3.10) for the $\boldsymbol{w}_i$ and fail to reject H_0 if $\boldsymbol{0}$ is in the prediction region (if $D_{\boldsymbol{0}} \le D_{(U_B)}$), reject H_0 otherwise.

The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is

$$\{\boldsymbol{w} : (\boldsymbol{w} - T)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \le D^2_{(U_B,T)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{(U_B,T)}\} \tag{3.11}$$

where $D^2_{(U_B,T)}$ is computed from $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$.

The Pelawa Watagoda and Olive (2018) hybrid large sample $100(1-\delta)\%$ confidence region shifts the hyperellipsoid (3.10) to be centered at T instead of $\overline{T}^*$:

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \le D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\}. \tag{3.12}$$

Hyperellipsoids (3.10) and (3.12) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (3.10) and (3.11) is

$$\left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g. \tag{3.13}$$

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_0 : \boldsymbol{\theta} \ne \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. For example, let $\boldsymbol{A}$ be a $g \times p$ matrix with full rank g, $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$, $\boldsymbol{\theta}_0 = \boldsymbol{0}$, and $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$. This section gives some theory for the bagging estimator $\overline{T}^*$, also called the smoothed bootstrap estimator. The theory may be useful for hypothesis testing after model selection if n/p is large. Empirically, bootstrapping with the bagging estimator often outperforms bootstrapping with T_n. See Efron (2014). See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator.

If i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, then under regularity conditions, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and v) $n\boldsymbol{S}_T^* \xrightarrow{P} Cov(\boldsymbol{u})$.

Suppose i) and ii) hold with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma_u}$. With respect to the bootstrap sample, T_n is a constant and the $\sqrt{n}(T_i^* - T_n)$ are iid for $i = 1, ..., B$. Let $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{v}_i \sim \boldsymbol{u}$ where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. Fix B. Then the average of the $\sqrt{n}(T_i^* - T_n)$ is

$$\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma_u}}{B}\right)$$

where $\boldsymbol{z} \sim AN_g(\boldsymbol{0}, \boldsymbol{\Sigma})$ is an asymptotic multivariate normal approximation. Hence as $B \to \infty$, $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$, and iii) and iv) hold. If B is fixed and $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma_u})$, then

$$\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim N_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma_u}}{B}\right) \text{ and } \sqrt{\mathrm{B}}\sqrt{\mathrm{n}}(\overline{\mathrm{T}}^* - \mathrm{T_n}) \xrightarrow{\mathrm{D}} \mathrm{N_g}(\boldsymbol{0}, \boldsymbol{\Sigma_u}).$$

Hence the prediction region method gives a large sample confidence region for $\boldsymbol{\theta}$ provided that the sample percentile $\hat{D}^2_{1-\delta}$ of the $D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*)$ is a consistent estimator of the percentile $D^2_{n,1-\delta}$ of the random variable $D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)$ in that $\hat{D}^2_{1-\delta} - D^2_{n,1-\delta} \xrightarrow{P} 0$. Since iii) and iv) hold, the sample percentile will be consistent under much weaker conditions than v) if $\boldsymbol{\Sigma_u}$ is nonsingular. For example, if $(n\boldsymbol{S}_T^*)^{-1} = \boldsymbol{\Sigma_u}^{-1} + \boldsymbol{C} + o_p(1)$ for some $g \times g$ constant matrix $\boldsymbol{C}$. Olive (2017b $\oint$ 5.3.3) proved that the prediction region method gives a large sample confidence region under the much stronger conditions of v) and $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma_u})$, but the above proof is simpler.

Now suppose that T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$ where $\sum_j \pi_{jn} = 1$, $\pi_{jn} \to \pi_j$ as $n \to \infty$, and $\sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j$ with $E(\boldsymbol{u}_j) = \boldsymbol{0}$

and $Cov(\boldsymbol{u}_j) = \boldsymbol{\Sigma}_j$. Then the cumulative distribution function (cdf) of T_n is $F_{T_n}(\boldsymbol{z}) = \sum_j \pi_{jn} F_{T_{jn}}(\boldsymbol{z})$ where $F_{T_{jn}}(\boldsymbol{z})$ is the cdf of T_{jn}. Hence

$$\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{3.14}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ and $F_{\boldsymbol{u}_j}(\boldsymbol{z})$ is the cdf of $\boldsymbol{u}_j$. Thus $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{\Sigma}_j$.

For the bootstrap, suppose that T_i^* is equal to T_{ij}^* with probability ρ_{jn} for $j = 1, ..., J$ where $\sum_j \rho_{jn} = 1$, and $\rho_{jn} \to \pi_j$ as $n \to \infty$. Let B_{jn} count the number of times $T_i^* = T_{ij}^*$ in the bootstrap sample. Then the bootstrap sample $T_1^*, ..., T_B^*$ can be written as

$$T_{1,1}^*, ..., T_{B_{1n},1}^*, ..., T_{1,J}^*, ..., T_{B_{Jn},J}^*$$

where the B_{jn} follow a multinomial distribution and $B_{jn}/B \xrightarrow{P} \rho_{jn}$ as $B \to \infty$. Conditionally on the B_{jn} and with respect to the bootstrap sample, the T_{ij}^* are independent. Denote $T_{1j}^*, ..., T_{B_{jn},j}^*$ as the jth bootstrap component of the bootstrap sample with sample mean $\overline{T}_j^*$ and sample covariance matrix $\boldsymbol{S}_{T,j}^*$. Then

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* = \sum_j \frac{B_{jn}}{B}\frac{1}{B_{jn}}\sum_{i=1}^{B_{jn}} T_{ij}^* = \sum_j \hat{\rho}_{jn}\overline{T}_j^*.$$

Suppose $\sqrt{n}(T_i^* - E(T^*)) \xrightarrow{D} \boldsymbol{v}_i \sim \boldsymbol{v}$ where $E(\boldsymbol{v}) = \boldsymbol{0}$, $Cov(\boldsymbol{v}) = \boldsymbol{\Sigma}_{\boldsymbol{v}}$, and $E(T^*) = \sum_j \rho_{jn} E(T_{ij}^*)$ where often $E(T_{ij}^*) = T_{jn}$. With respect to the data distribution, suppose $\sqrt{n}(E(T^*) - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$. Then by an argument similar to the one given for when T_n is not from a mixture distribution, $\sqrt{n}(\overline{T}^* - E(T^*)) \xrightarrow{P} \boldsymbol{0}$, $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{v}$, and $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$.

Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \le \hat{D}^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \le \hat{D}^2_{(U_B)}\}$ centered at a randomly selected T_n contains $\overline{T}$ with probability $1 - \delta_B$. If i) holds with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$, then for fixed B,

$$\sqrt{n}(\overline{T} - \boldsymbol{\theta}) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right).$$

Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. Hence R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. We also need $(n\boldsymbol{S}_T)^{-1}$ to be "not too ill conditioned."

With a mixture distribution, the bootstrap sample shifts the data cloud to be centered at $\overline{T}^*$ where $\sqrt{n}(\overline{T}^* - \sum_j \rho_{jn} T_{jn}) \xrightarrow{P} \mathbf{0}$. The T_{jn} are computed from the same data set and hence correlated. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$, and $(n\boldsymbol{S}^*_T)^{-1}$ is "not too ill conditioned." Then

$$D_1^2 = D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}^*_T) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}^*_T)^{-1} \sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D^2_{\boldsymbol{\theta}}(T_n, \boldsymbol{S}^*_T) = \sqrt{n}(T_n - \boldsymbol{\theta})^T (n\boldsymbol{S}^*_T)^{-1} \sqrt{n}(T_n - \boldsymbol{\theta}), \quad \text{and}$$

$$D_3^2 = D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}^*_T) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T (n\boldsymbol{S}^*_T)^{-1} \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})$$

are well behaved in that there exist cutoffs $\hat{D}^2_{i,1-\delta}$ that would result in good confidence regions for $i = 2$ and 3. Heuristically, for a mixture distribution, the deviation $\overline{T}^* - \boldsymbol{\theta}$ tends to be smaller on average than the deviations $T_n - \boldsymbol{\theta} \approx T_i^* - \overline{T}^*$, while the deviation $T_i^* - T_n$ tends to be larger than the other three deviations, on average. Hence $\hat{D}^2_{2,1-\delta} = D^2_{(U_B)}$ gives coverage close to the nominal coverage for prediction region (3.12), but cutoffs $\hat{D}^2_{3,1-\delta} = D^2_{(U_B)}$ and $\hat{D}^2_{2,1-\delta} = D^2_{(U_B,T)}$ are slightly too large, and prediction regions (3.10) and (3.11) tend to have coverage slightly higher than the nominal coverage $1-\delta$ if n and B are large. In simulations for $n \geq 20p$, the coverage tends to get close to $1-\delta$ for $B \geq \max(400, 50p)$ so that $\boldsymbol{S}^*_T$ is a good estimator of $Cov(T^*)$.

To examine the bagging estimator, assume that each bootstrap component satisfies vi) $\sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_j)$, vii) $\sqrt{n}(T^*_{ij} - T_{jn}) \xrightarrow{D} \boldsymbol{u}_j$, viii) $\sqrt{n}(\overline{T}^*_j - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j$, ix) $\sqrt{n}(T^*_{ij} - \overline{T}^*_j) \xrightarrow{D} \boldsymbol{u}_j$, x) $n\boldsymbol{S}^*_{T,j} \xrightarrow{P} \boldsymbol{\Sigma}_j$, and xi) $\sqrt{n}(T_{jn} - \overline{T}^*_j) \xrightarrow{P} \mathbf{0}$ as $B_{jn} \to \infty$ and $n \to \infty$.

Consider the random vectors

$$Z_n = \sum_j \frac{B_{jn}}{B} T_{jn} \quad \text{and} \quad \mathrm{W_n} = \sum_{\mathrm{j}} \rho_{\mathrm{jn}} \mathrm{T_{jn}}.$$

By xi)

$$\sqrt{n}(Z_n - \overline{T}^*) = \sqrt{n}(\sum_j \frac{B_{jn}}{B} T_{jn} - \overline{T}^*) = \sum_j \frac{B_{jn}}{B} \sqrt{n}(T_{jn} - \overline{T}^*_j) \xrightarrow{P} \mathbf{0}.$$

Also, $\sqrt{n}(Z_n - \boldsymbol{\theta}) - \sqrt{n}(W_n - \boldsymbol{\theta}) =$

$$\sum_j \left(\frac{B_{jn}}{B} - \rho_{jn} \right) \sqrt{n}(T_{jn} - \boldsymbol{\theta}) = \sum_j O_P(1) O_P(n^{-1/2}) \xrightarrow{P} \mathbf{0}.$$

Assume the $\boldsymbol{u}_{nj} = \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j$ are such that

$$\sqrt{n}(W_n - \boldsymbol{\theta}) = \sum_j \rho_{jn} \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w} = \sum_j \pi_j \boldsymbol{u}_j.$$

Note that $E(\boldsymbol{w}) = \mathbf{0}$ and $Cov(\boldsymbol{w}) = \boldsymbol{\Sigma}_{\boldsymbol{w}} = \sum_j \sum_k \pi_j \pi_k Cov(\boldsymbol{u}_j, \boldsymbol{u}_k)$. Hence

$$\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}. \tag{3.15}$$

Since $\boldsymbol{w}$ is a weighted mean of the $\boldsymbol{u}_j \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_j)$, a normal approximation is $\boldsymbol{w} \approx N_g(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$. The approximation is exact if the $\boldsymbol{u}_j$ with positive π_j have a joint multivariate normal distribution.

Now consider variable selection for model (1.1) with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$. Olive (2017a: p. 128, 2018a) showed that the prediction region method can simulate well for the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Assume p is fixed, $n \geq 20p$, and that the error distribution is unimodal and not highly skewed. The response plot and residual plot are plots with $\hat{Y} = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$ on the horizontal axis and Y or r on the vertical axis, respectively. Then the plotted points in these plots should scatter in roughly even bands about the identity line (with unit slope and zero intercept) and the $r = 0$ line, respectively. If the error distribution is skewed or multimodal, then much larger sample sizes may be needed.

For the nonparametric bootstrap, cases are sampled with replacement, and the above conditions hold since each component bootstraps correctly. For the residual bootstrap, we use the fitted values and residuals from the OLS full model, but fit $\hat{\boldsymbol{\beta}}$ for a method such as forward selection, lasso, et cetera. Consider forward selection where each component uses a $\hat{\boldsymbol{\beta}}_{I_j}$. Let $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ be the fitted values from the OLS full model where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $\boldsymbol{r}^W$ denote an $n \times 1$ random vector of elements selected with replacement from the OLS full model residuals. Following Freedman (1981) and Efron (1982, p. 36), $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ follows a standard linear model where the elements r_i^W of $\boldsymbol{r}^W$ are iid from the empirical distribution of the OLS full model residuals r_i. Hence

$$E(r_i^W) = \frac{1}{n}\sum_{i=1}^n r_i = 0, \quad V(r_i^W) = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^n r_i^2 = \frac{n-p}{n} MSE,$$

$$E(\boldsymbol{r}^W) = \boldsymbol{0}, \text{ and } \operatorname{Cov}(\boldsymbol{Y}^*) = \operatorname{Cov}(\boldsymbol{r}^{\mathrm{W}}) = \sigma_{\mathrm{n}}^2 \boldsymbol{I}_{\mathrm{n}}.$$

Then $\hat{\boldsymbol{\beta}}^*_{I_j} = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{Y}^* = \boldsymbol{D}_j\boldsymbol{Y}^*$ with $Cov(\hat{\boldsymbol{\beta}}^*_{I_j}) = \sigma_n^2(\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}$ and $E(\hat{\boldsymbol{\beta}}^*_{I_j}) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{I_j}$ since $\boldsymbol{H}\boldsymbol{X}_{I_j} = \boldsymbol{X}_{I_j}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

For the above residual bootstrap with forward selection and C_p, let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and $T_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0} = \boldsymbol{A}\boldsymbol{D}_{j,0}\boldsymbol{Y}$ where $\boldsymbol{D}_{j,0}$ adds rows of zeroes to $\boldsymbol{D}_j$ corresponding to the x_i not in I_j. If $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \sigma^2\boldsymbol{V}_j)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}_{j,0})$ where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeroes corresponding to the x_i not in I_j. Then under regularity conditions, (3.14) and (3.15) hold where $\sqrt{n}(\sum_j \rho_{jn}T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$, and the sum is over $j : S \subseteq I_j$. Thus $E(T^*) = \sum_j \rho_{jn}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$ and $\boldsymbol{S}_T^*$ is a consistent estimator of $Cov(T^*)$

$$= \sum_j \rho_{jn} Cov(T_{jn}^*) + \sum_j \rho_{jn}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}\hat{\boldsymbol{\beta}}_{I_j,0}^T\boldsymbol{A}^T - E(T^*)[E(T^*)]^T$$

where asymptotically the sum is over $j : S \subseteq I_j$. If $\boldsymbol{\theta}_0 = \boldsymbol{0}$, then $n\boldsymbol{S}_T^* = \boldsymbol{\Sigma}_A + O_P(1)$ where

$$nCov(T_n) \xrightarrow{P} \boldsymbol{\Sigma}_A = \sum_j \sigma^2\pi_j\boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T.$$

Then $(n\boldsymbol{S}_T^*)^{-1}$ tends to be "well behaved" if $\boldsymbol{\Sigma}_A$ is nonsingular. The prediction region (3.10) bootstraps T_n, but uses $\overline{T}^*$ to increase the coverage for moderate samples.

Some special cases are also interesting. Suppose $\pi_d = 1$ so $\boldsymbol{u} \sim \boldsymbol{u}_d \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_d)$. This occurs for C_p if $a_S = p$ so S is the full model, and for methods like BIC that choose I_S with probability going to one. Knight and Fu (2000) had similar bootstrap results for this case. Next, if for each $\pi_j > 0$, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}_j\boldsymbol{A}^T) = N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$, then $\boldsymbol{A}\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

In the simulations where S is not the full model, inference with forward selection with I_{min} using C_p appears to be more precise than inference with the OLS full model if $n \geq 20p$ and $B \geq 50p$. Higher than nominal coverage can occur because of the zero padding. It is possible that $\boldsymbol{S}_T^*$ is singular if a column of the bootstrap sample is equal to $\boldsymbol{0}$.

Examining $\hat{\boldsymbol{\beta}}_S$ and $\hat{\boldsymbol{\beta}}_E$ is informative for I_{min}. See Equation (1.3). First assume that the nontrivial predictors are orthogonal or uncorrelated with zero mean so $\boldsymbol{X}^T\boldsymbol{X}/n \to diag(d_1, ..., d_p)$ as $n \to \infty$ where each $d_i > 0$. Then $\hat{\boldsymbol{\beta}}_S$ has the same multivariate normal limiting distribution for I_{min} and for the OLS full model. The bootstrap distribution for $\hat{\boldsymbol{\beta}}_E$ is a mixture of zeros and a distribution that would produce a confidence region for $\boldsymbol{A}\boldsymbol{\beta}_E = \boldsymbol{0}$ that has asymptotic coverage of $\boldsymbol{0}$ equal to $100(1-\delta)\%$. Hence the asymptotic coverage is greater than the nominal coverage provided that $\boldsymbol{S}_T^*$ in nonsingular with probability going to one (e.g., $p - a_S$ is small), where $T = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{E,I_{min},0}$. For uncorrelated predictors with zero mean, the number of bootstrap samples $B \geq 50p$ may work well for the shorth confidence intervals and for testing $\boldsymbol{A}\boldsymbol{\beta}_S = \boldsymbol{0}$.

In the simulations for forward selection, coverages did not change much as the ρ was increased from zero to near one, where ρ was the correlation between any two nontrivial predictors. Under model (1.3), we still have that $\hat{\boldsymbol{\beta}}_{I_j,0}$ is a $\sqrt{n}$ consistent asymptotically normal estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_S^T, \boldsymbol{\beta}_E^T)^T$ where $\boldsymbol{\beta}_E = \boldsymbol{0}$. Hence the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta})$ is a mixture of $N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{V}_{j,0})$ distributions, and the limiting distribution of $\sqrt{n}(\hat{\beta}_{i,I_{min},0} - \beta_i)$ is a mixture of $N(0, \sigma^2_{ij})$ distributions. For a β_i that is a component of $\boldsymbol{\beta}_S$, the symmetric mixture distribution has a pdf. Then the simulated shorth confidence intervals have coverage near the nominal coverage if n and B are large enough.

Note that there are several important variable selection models, including the model given by Equation (1.3). Another model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T \boldsymbol{\beta}_{S_i}$ for $i = 1, ..., J$. Then there are $J \geq 2$ competing "true" nonnested submodels where $\boldsymbol{\beta}_{S_i}$ is $a_{S_i} \times 1$. For example, suppose the $J = 2$ models have predictors x_1, x_2, x_3 for S_1 and x_1, x_2, x_4 for S_2. Then x_3 and x_4 are likely to be selected and omitted often by forward selection for the B bootstrap samples.

Hence omitting all predictors x_i that have a $\beta^*_{ij} = 0$ for at least one of the bootstrap samples $j = 1, ..., B$ could result in underfitting, e.g. using just x_1 and x_2 in the above $J = 2$ example. If n and B are large enough, the singleton set $\{\mathbf{0}\}$ could still be the "100%" confidence region for a vector $\boldsymbol{\beta}_O$.

Suppose the predictors x_i have been standardized. Then another important regression model has the β_i taper off rapidly, but no coefficients are equal to zero. For example, $\beta_i = e^{-i}$ for $i = 1, ..., p$.

Bootstrap Confidence Intervals

For $g = 1$, the percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T^*_i from a bootstrap sample $T^*_1, ..., T^*_B$ where the statistic T_n is an estimator of θ based on a sample of size n. Note that the squared Mahalanobis distance $D^2_\theta = (\theta - \overline{T^*})^2 / S^{2*}_T \leq D^2_{(U_B)}$ is equivalent to $\theta \in [\overline{T^*} - S^*_T D_{(U_B)}, \overline{T^*} + S^*_T D_{(U_B)}]$, which is an interval centered at $\overline{T^*}$ just long enough to cover U_B of the T^*_i. Hence the prediction region method is a special case of the percentile method if $g = 1$. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The Frey (2013) shorth(c) interval (3.1) (with c given by (3.3)) applied to the T^*_i is recommended since the shorth confidence interval can be much shorter than the Efron (2014) or prediction region method confidence intervals if $g = 1$. The shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples. Note that if $\sqrt{n}(T_n - \theta) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T^*_i - \theta) \xrightarrow{D} \boldsymbol{u}$ where $\boldsymbol{u}$ has a symmetric probability density function, then the shorth confidence interval is asymptotically equivalent to the usual percentile method confidence interval that uses the central proportion of the bootstrap sample.

Note that correction factors $b_n \to 1$ are used in large sample confidence intervals and tests if the limiting distribution is N(0,1) or χ^2_p, but a t_{d_n} or pF_{p,d_n} cutoff is used: $t_{d_n,1-\delta}/z_{1-\delta} \to 1$ and $pF_{p,d_n,1-\delta}/\chi^2_{p,1-\delta} \to 1$ if $d_n \to \infty$ as $n \to 1$. Using correction factors for prediction intervals and bootstrap confidence regions improves the performance for moderate sample size n.

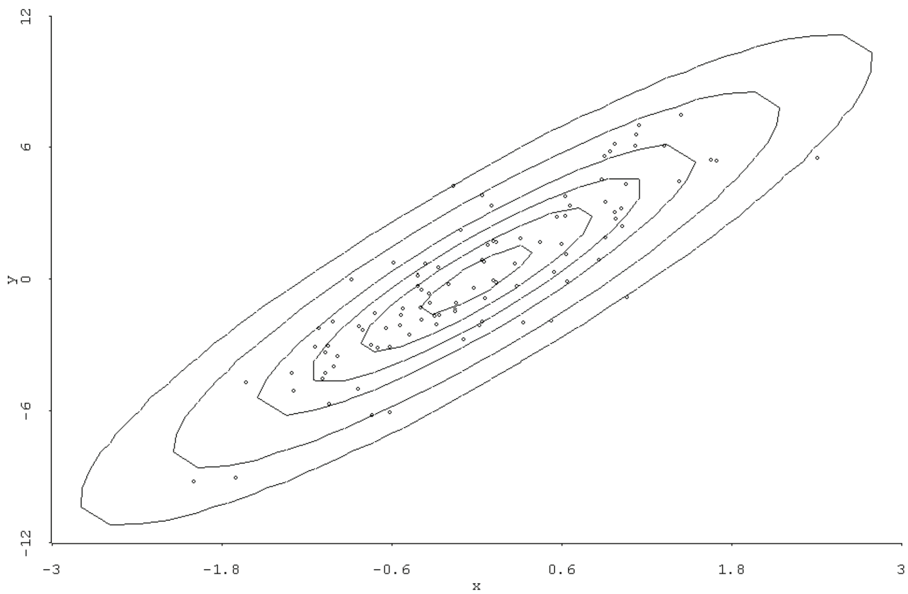
CHAPTER 4

EXAMPLE AND SIMULATIONS

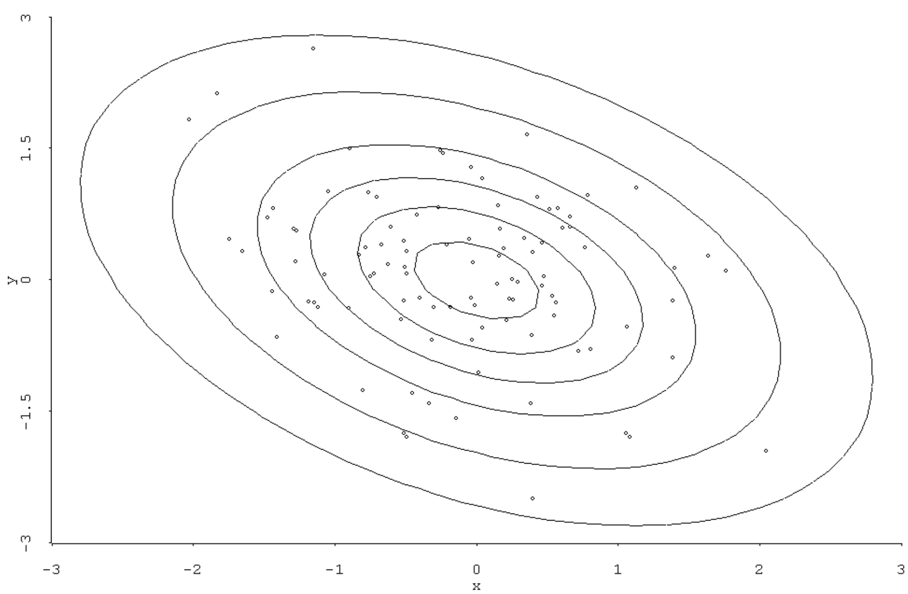
Figure 1 shows 10%, 30%, 50%, 70%, 90% and 98% prediction regions for a future value of T_f for two multivariate normal distributions. The plotted points are iid $T_1, ..., T_B$ with $B = 100$.

Example. The Hebbler (1847) data was collected from $n = 26$ districts in Prussia in 1843. We will study the relationship between $Y =$ the *number of women married to civilians* in the district with the predictors $x_1 =$ constant, $x_2 = pop =$ the *population of the district in 1843*, $x_3 = mmen =$ the *number of married civilian men* in the district, $x_4 = mmilmen$ $=$ *number of married men in the military* in the district, and $x_5 = milwmn =$ the *number of women married to husbands in the military* in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. Hence Y and X_3 are highly correlated but not equal. Similarly, x_4 and x_5 are highly correlated but not equal. We expect that $Y = x_3 + e$ is a good model. Forward selection with C_p selected the model a constant and *mmen*.

Let $\boldsymbol{x} = (1 \ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ where the $m = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $Cov(\boldsymbol{u}_i) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (m-2)\psi^2]$. Hence the correlations are $cor(x_i, x_j) = \rho = (2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let $Y_i = 1 + 1x_{i,2} + \cdots + 1x_{i,k+1} + e_i$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (1, .., 1, 0, ..., 0)^T$ with $k+1$ ones and $p-k-1$ zeros. The zero mean errors e_i were iid from five distributions: i) N(0,1), ii) t_3, iii) EXP(1) - 1, iv) uniform$(-1, 1)$, and v) 0.9 N(0,1) + 0.1 N(0,100). Only



)



b)

Figure 4.1. Prediction Regions

distribution iii) is not symmetric.

A small simulation was done using $B = \max(1000, n, 20p)$ and 5000 runs. So an observed coverage in [0.94, 0.96] gives no reason to doubt that the CI or confidence region has the nominal coverage of 0.95. The simulation used $p = 4, 6, 7, 8$, and 10; $n = 25p$, n=50p, $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and $p - 2$.

When $\psi = 0$, the full model least squares confidence intervals for β_i should have length near $2t_{96,0.975}\sigma/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when the iid zero mean errors have variance σ^2. The simulation computed the Frey shorth(c) interval for each β_i and used bootstrap confidence regions to test whether first $k+1$ $\beta_i = 1$ and the last $p-k-1$ $\beta_i = 0$. The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 would suggest coverage is close to the nominal value.

The regression models used the residual bootstrap on the forward selection estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Table 1 gives results for when the iid errors $e_i \sim N(0,1)$. Two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The last six columns give results for the tests. The length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ or $[0, D_{(U_B),T}]$ where $D_{(U_B)}$ or $D_{(U_B),T}$ is the cutoff for the confidence region. Volumes of the confidence regions can be compared using (3.13). The first two lines of the table correspond to the R output shown below, with $g = 2$.

```
library(leaps);Y <- marry[,3]; X <- marry[,-3]
temp<-regsubsets(X,Y,method="forward")
out<-summary(temp)
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000
Selection Algorithm: forward
         pop mmen mmilmen milwmn
1  ( 1 ) " " "*"  " "     " "
2  ( 1 ) " " "*"  "*"     " "
```

```
3  ( 1 ) "*" "*"  "*"     " "
4  ( 1 ) "*" "*"  "*"     "*"
record coverages and ``lengths" for
b1, b2, bp-1, bp, pm0,  hyb0,   BR0,  pm1,   hyb1,  BR1,
library(leaps)
vsbootsim4(n=100,p=4,k=1,nruns=5000,type=1,psi=0)
$cicov
 [1] 0.9480 0.9496 0.9972 0.9958 0.9910 0.9786 0.9914 0.9384 0.9394 1.0000
$avelen
 [1] 0.3954381 0.3987018 0.3232973 0.3231127 2.6987198 2.6987198 3.0020469
 [8] 2.6987198 2.6987198 3.0020469
$beta
[1] 1 1 0 0
$k
[1] 1
```

Table 4.1. Bootstrapping OLS Forward Selection with C_p, $e_i \sim N(0,1)$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
100,4,1,0	0.9440	0.9410	0.9978	0.9978	0.9936	0.9808	0.9944	0.9326	0.9342	0.9344
len	0.3960	0.3990	0.3245	0.3239	2.6898	2.6898	2.9960	2.4499	2.4499	2.4568
100,4,1,0.5	0.9456	0.9696	0.9982	0.9974	0.9924	0.9842	0.9936	0.9532	0.9548	0.9602
len	0.3958	0.65801	0.5359	0.5390	2.6968	2.6968	2.9878	2.4584	2.4584	2.5763
100,4,1,0.9	0.9436	0.9674	0.9962	0.9976	0.9926	0.9836	0.9950	0.9608	0.9570	0.9656
len	0.3966	2.7667	2.7378	2.7336	2.7175	2.7175	2.9711	2.4990	2.4990	2.6012
100,4,2,0	0.9492	0.9430	0.9410	0.9982	0.9976	0.9342	0.9970	0.9364	0.9366	0.9364
len	0.3958	0.3989	0.3986	0.3240	2.1371	2.1371	2.3933	2.7999	2.7999	2.8044
100,4,2,0.5	0.9424	0.9542	0.9510	0.9974	0.9962	0.9326	0.9958	0.9570	0.9572	0.9640
len	0.3963	0.6595	0.6591	0.5366	2.1393	2.1393	2.3990	2.8477	2.8477	2.9635
100,4,2,0.9	0.9454	0.9224	0.9262	0.9978	0.9970	0.8770	0.9980	0.9800	0.9704	0.9836
len	0.3965	2.70245	2.7045	2.6508	2.0815	2.0815	2.4272	2.9110	2.9110	3.1614
300,6,1,0	0.9496	0.9460	0.9972	0.9974	0.9938	0.9938	0.9966	0.9428	0.9424	0.9428
len	0.2230	0.23055	0.1864	0.1873	3.3731	3.3731	3.7299	2.4503	2.4503	2.4550
300,6,1,0.4082	0.9486	0.9634	0.9990	0.9992	0.9946	0.9916	0.9960	0.9540	0.9560	0.9608
len	0.2301	0.3524	0.2848	0.2856	3.3786	3.3786	3.7162	2.4481	2.4481	2.5314
300,6,1,0.9	0.9470	0.9796	0.9992	0.9980	0.9934	0.9902	0.9962	0.9464	0.9320	0.9558
len	0.2299	1.8921	1.6855	1.6924	3.3400	3.3400	3.7184	2.4309	2.4309	2.5538
300,6,4,0	0.9466	0.9534	0.9480	0.9978	0.9970	0.9386	0.9962	0.9440	0.9430	0.9436
len	0.2298	0.2305	0.2305	0.1856	2.1381	2.1381	2.3944	3.3383	3.3383	3.3412
300,6,4,0.4082	0.9506	0.9516	0.9480	0.9982	0.9974	0.9398	0.9970	0.9564	0.9568	0.9630
len	0.2302	0.3502	0.3502	0.2841	2.1473	2.1473	2.3981	3.3687	3.3687	3.4470
300,6,4,0.9	0.9504	0.9488	0.9538	0.9980	0.9966	0.9336	0.9972	0.9548	0.9038	0.9546
len	0.2304	1.9736	1.9746	1.7547	2.1201	2.1201	2.3596	3.4077	3.4077	3.6859

Table 4.2. Bootstrapping OLS Forward Selection with C_p, $e_i \sim N(0,1)$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
175,7,1,0	0.9464	0.9518	0.9978	0.9970	0.9926	0.9916	0.9956	0.9412	0.9420	0.9444
len	0.3005	0.3018	0.2447	0.2450	3.6259	3.6259	4.0158	2.4503	2.4503	2.4599
175,7,1,0.3780	0.9424	0.9584	0.9966	0.9984	0.9916	0.9904	0.9950	0.9462	0.9438	0.9500
len	0.3007	0.4475	0.3638	0.3645	3.6256	3.6256	3.9925	2.4486	2.4486	2.5243
175,7,1,0.9	0.9480	0.9608	0.9982	0.9966	0.9940	0.9932	0.9974	0.9602	0.9558	0.9670
len	0.3003	2.4054	2.2929	2.2798	3.6600	3.6600	4.0338	2.4661	2.4661	2.6104
175,7,5,0	0.9428	0.9510	0.9448	0.9962	0.9950	0.9334	0.9946	0.9322	0.9322	0.9332
len	0.3004	0.3014	0.3018	0.2441	2.1347	2.1347	2.3930	3.5631	3.5631	3.5675
175,7,5,0.3780	0.9478	0.9440	0.9462	0.9980	0.9972	0.9344	0.9972	0.9466	0.9478	0.9536
len	0.3010	0.4453	0.4453	0.3595	2.1313	2.1313	2.3917	3.5851	3.5851	3.6558
175,7,5,0.9	0.9448	0.9374	0.9334	0.9964	0.9942	0.9170	0.9940	0.9634	0.9176	0.9686
len	0.3017	2.4924	2.4812	2.3226	2.1185	2.1185	2.3659	3.6605	3.6605	4.0084
400,8,1,0	0.9538	0.9480	0.9988	0.9982	0.9974	0.9976	0.9984	0.9464	0.9478	0.9496
len	0.1992	0.1998	0.1604	0.1614	3.8692	3.8692	4.2738	2.4494	2.4494	2.4552
400,8,1,0.3536	0.9500	0.9598	0.9990	0.9986	0.9958	0.9950	0.9982	0.9544	0.9550	0.9616
len	0.1992	0.2887	0.2337	0.2342	3.8698	3.8698	4.2517	2.4484	2.4484	2.5167
400,8,1,0.9	0.9562	0.9796	0.9990	0.9986	0.9966	0.9946	0.9986	0.9474	0.9214	0.9582
len	0.1993	1.7524	1.5142	1.5175	3.8375	3.8375	4.2488	2.4350	2.4350	2.5507
400,8,6,0	0.9442	0.9486	0.9494	0.9978	0.9974	0.9392	0.9970	0.9364	0.9362	0.9366
len	0.1994	0.1999	0.1996	0.1613	2.1324	2.1324	2.3891	3.7691	3.7691	3.7717
400,8,6,0.3536	0.9422	0.9500	0.9546	0.9976	0.9972	0.9356	0.9968	0.9508	0.9524	0.9568
len	0.1991	0.2864	0.2869	0.2319	2.1475	2.1475	2.3978	3.7904	3.7904	3.8504
400,8,6,0.9	0.9492	0.9504	0.9484	0.9966	0.9952	0.9282	0.9950	0.9312	0.8432	0.9180
len	0.1999	1.7977	1.8116	1.5520	2.1298	2.1298	2.3708	3.8348	3.8348	4.1058

Table 4.3. Bootstrapping OLS Forward Selection with C_p, $e_i \sim N(0,1)$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
250,10,1,0	0.9500	0.9456	0.9976	0.9980	0.9968	0.9974	0.9988	0.9430	0.9428	0.9450
len	0.2521	0.2527	0.2048	0.2049	4.2704	4.2704	4.7180	2.4502	2.4502	2.4606
250,10,1,0.3162	0.9416	0.9550	0.9986	0.9980	0.9966	0.9968	0.9988	0.9460	0.9482	0.9530
len	0.2517	0.3504	0.2843	0.2848	4.2645	4.2645	4.6965	2.4492	2.4492	2.5113
250,10,1,0.9	0.9476	0.9582	0.9982	0.9976	0.9960	0.9962	0.9990	0.9506	0.9440	0.9622
len	0.2514	2.1333	1.9590	1.9532	4.2714	4.2714	4.7180	2.4516	2.4516	2.5983
250,10,8,0.9	0.9478	0.9462	0.9426	0.9978	0.9970	0.9304	0.9958	0.9486	0.8430	0.9354
len	0.2529	2.1942	2.1965	1.9906	2.1259	2.1259	2.3694	4.2105	4.2105	4.5796

Table 4.4. Bootstrapping OLS Forward Selection with C_p, $e_i \sim t_3$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
100,4,1,0	0.9436	0.9510	0.9966	0.9978	0.9932	0.9804	0.9924	0.9470	0.9506	0.9516
len	0.6558	0.6706	0.5388	0.5415	2.7146	2.7146	3.0198	2.4747	2.4747	2.4821
100,4,1,0.5	0.9510	0.9542	0.9978	0.9974	0.9916	0.9736	0.9932	0.9612	0.9538	0.9648
len	0.6591	1.1189	0.9315	0.9336	2.7331	2.7331	3.0244	2.4953	2.4953	2.6200
100,4,1,0.9	0.9392	0.9468	0.9952	0.9950	0.9892	0.9796	0.9938	0.9660	0.9646	0.9712
len	0.6579	4.6420	4.6574	4.6365	2.7875	2.7875	3.0313	2.5666	2.5666	2.6915
100,4,2,0	0.9426	0.9450	0.9434	0.9970	0.9950	0.9304	0.9952	0.9450	0.9446	0.9460
len	0.6561	0.6739	0.6730	0.5419	2.1364	2.1364	2.3967	2.8554	2.8554	2.8624
100,4,2,0.5	0.9442	0.9428	0.9422	0.9984	0.9966	0.9314	0.9960	0.9638	0.9502	0.9628
len	0.6611	1.1456	1.1469	0.9434	2.1269	2.1269	2.3821	2.9254	2.9254	3.0497
100,4,2,0.9	0.9432	0.9024	0.9066	0.9976	0.9972	0.9328	0.9966	0.9836	0.9776	0.9870
len	0.6576	4.4685	4.4695	4.4649	1.9951	1.9951	2.1941	2.9722	2.9722	3.1843
300,6,1,0	0.9520	0.9510	0.9982	0.9984	0.9962	0.9944	0.9974	0.9488	0.9476	0.9474
len	0.3867	0.3904	0.3156	0.3143	3.4111	3.4111	3.7595	2.4657	2.4657	2.4703
300,6,1,0.4082	0.9484	0.9612	0.9986	0.9984	0.9954	0.9930	0.9972	0.9536	0.9558	0.9608
len	0.3878	0.5998	0.4836	0.4837	3.4170	3.4170	3.7510	2.4639	2.4639	2.5441
300,6,1,0.9	0.9456	0.9496	0.9984	0.9988	0.9970	0.9958	0.9982	0.9684	0.9668	0.9738
len	0.3885	3.0089	2.9342	2.9292	3.4516	3.4516	3.7801	2.5109	2.51092	2.6502
300,6,4,0	0.9424	0.9540	0.9496	0.9984	0.9980	0.9390	0.9978	0.9450	0.9446	0.9450
len	0.3884	0.3926	0.3929	0.3148	2.1431	2.1431	2.3983	3.4017	3.4017	3.4053
300,6,4,0.4082	0.9454	0.9508	0.9496	0.9984	0.9972	0.9304	0.9964	0.9638	0.9632	0.9694
len	0.3901	0.6014	0.6010	0.4847	2.1429	2.1429	2.4019	3.4343	3.4343	3.5109
300,6,4,0.9	0.9550	0.9248	0.9246	0.9978	0.9962	0.9230	0.9962	0.9836	0.9722	0.9888
len	0.3876	3.0558	3.0582	2.9237	2.1195	2.1195	2.3626	3.5045	3.5045	3.8389

Table 4.5. Bootstrapping OLS Forward Selection with C_p, $e_i \sim t_3$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
175,7,1,0	0.9510	0.9494	0.9974	0.9974	0.9966	0.9946	0.9974	0.9452	0.9478	0.9486
len	0.5039	0.5116	0.4114	0.4140	3.6891	3.6891	4.0617	2.4698	2.4698	2.4787
175,7,1,0.3780	0.9486	0.9582	0.9982	0.9992	0.9962	0.9946	0.9974	0.9586	0.9566	0.9614
len	0.5057	0.7627	0.6149	0.6109	3.6875	3.6875	4.0464	2.4704	2.470	2.5471
175,7,1,0.9	0.9454	0.9018	0.9980	0.9982	0.9946	0.9942	0.9976	0.9700	0.9656	0.9756
len	0.5027	3.9474	3.9135	3.9183	3.6972	3.6972	4.0379	2.5339	2.5339	2.6911
175,7,5,0	0.9452	0.9492	0.9464	0.9984	0.9972	0.9350	0.9968	0.9442	0.9446	0.9450
len	0.5043	0.5130	0.5116	0.4112	2.1404	2.1404	2.4023	3.6584	3.6584	3.6634
175,7,5,0.3780	0.9506	0.9470	0.9470	0.9986	0.9976	0.9338	0.9974	0.9646	0.9652	0.9708
len	0.5051	0.7612	0.7618	0.6125	2.1466	2.1466	2.4086	3.7057	3.7057	3.7750
175,7,5,0.9	0.9464	0.8958	0.8904	0.9984	0.9978	0.9250	0.9956	0.9900	0.9816	0.9930
len	0.5067	3.9892	3.9786	3.8957	2.1262	2.1262	2.3700	3.8004	3.8004	4.158
400,8,1,0	0.9494	0.9538	0.9976	0.9988	0.9956	0.9954	0.9974	0.9482	0.9484	0.9482
len	0.3359	0.3393	0.2725	0.2742	3.9153	3.9153	4.3088	2.4649	2.4649	2.4703
400,8,1,0.3536	0.9540	0.9598	0.9986	0.9978	0.9964	0.9958	0.9982	0.9542	0.9552	0.9590
len	0.3384	0.4938	0.3971	0.4006	3.9177	3.9177	4.2890	2.4638	2.4638	2.5276
400,8,1,0.9	0.9516	0.9440	0.9984	0.9986	0.9966	0.9954	0.9988	0.9682	0.9632	0.9750
len	0.3383	2.7223	2.6108	2.6050	3.9369	3.9369	4.3109	2.4971	2.4971	2.6486
400,8,6,0	0.9494	0.9488	0.9546	0.9976	0.9972	0.9332	0.9974	0.9454	0.9466	0.9468
len	0.3384	0.3417	0.3421	0.2738	2.1403	2.1403	2.3931	3.8510	3.8510	3.8544
400,8,6,0.3536	0.9494	0.9512	0.9500	0.9980	0.9970	0.9342	0.9972	0.9588	0.9590	0.9636
len	0.3382	0.4906	0.4902	0.3947	2.1440	2.1440	2.3963	3.8711	3.8711	3.9292
400,8,6,0.9	0.9482	0.9256	0.9254	0.9982	0.9968	0.9258	0.9970	0.9796	0.9484	0.9826
len	0.3388	2.7974	2.7759	2.6302	2.1323	2.1323	2.3705	3.9454	3.9454	4.3042

Table 4.6. Bootstrapping OLS Forward Selection with C_p, $e_i \sim t_3$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
250,10,1,0	0.9426	0.9444	0.9984	0.9988	0.9980	0.9976	0.9994	0.9406	0.9376	0.9396
len	0.4211	0.4260	0.3455	0.3409	4.3450	4.3450	4.7780	2.4676	2.4676	2.4778
250,10,1,0.3162	0.9456	0.9574	0.9986	0.9982	0.9972	0.9976	0.9988	0.9512	0.9542	0.9598
len	0.4230	0.5950	0.4779	0.4784	4.3421	4.3421	4.7582	2.4664	2.4664	2.5242
250,10,1,0.9	0.9484	0.8972	0.9990	0.9974	0.9970	0.9974	0.9992	0.9660	0.9650	0.9718
len	0.4265	3.4347	3.3476	3.3511	4.3515	4.3515	4.7672	2.5210	2.5210	2.6866
250,10,8,0.9	0.9416	0.9042	0.9062	0.9982	0.9962	0.9220	0.9964	0.9836	0.9614	0.9874
len	0.4243	3.4498	3.4512	3.3359	2.1203	2.1203	2.3622	4.3710	4.3710	4.7840

Table 4.7. Bootstrapping OLS Forward Selection with C_p, $e_i \sim EXP(1) - 1$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
100,4,1,0	0.9370	0.9454	0.9984	0.9982	0.9956	0.9820	0.9950	0.9294	0.9278	0.9284
len	0.3915	0.3971	0.3234	0.3217	2.7116	2.7116	3.0153	2.4542	2.4542	2.4618
100,4,1,0.5	0.9450	0.9712	0.9978	0.9976	0.9920	0.9852	0.9948	0.9532	0.9542	0.9634
len	0.3917	0.6584	0.5340	0.5331	2.7141	2.7141	3.0081	2.4616	2.4616	2.5801
100,4,1,0.9	0.9400	0.9668	0.9960	0.9962	0.9922	0.9830	0.9940	0.9530	0.9486	0.9592
len	0.3917	2.7537	2.7166	2.7131	2.7099	2.7099	2.9546	2.4934	2.4934	2.6020
100,4,2,0	0.9376	0.9464	0.9506	0.9978	0.9974	0.9346	0.9966	0.9290	0.9286	0.9298
len	0.3941	0.4004	0.4002	0.3243	2.1434	2.1434	2.4011	2.8194	2.8194	2.8241
100,4,2,0.5	0.9386	0.9530	0.9580	0.9984	0.9976	0.9326	0.9976	0.9570	0.9572	0.9634
len	0.3923	0.6608	0.6596	0.5368	2.1325	2.1325	2.3941	2.8622	2.8622	2.9767
100,4,2,0.9	0.9318	0.9318	0.9252	0.9984	0.9978	0.8894	0.9982	0.9698	0.9600	0.9760
len	0.3930	2.7059	2.7001	2.6587	2.0913	2.0913	2.4156	2.9165	2.9165	3.1686
300,6,1,0	0.9484	0.9456	0.9978	0.9984	0.9958	0.9956	0.9978	0.9416	0.9440	0.9442
len	0.2290	0.2301	0.1863	0.1856	3.3927	3.3927	3.7456	2.4521	2.4521	2.4568
300,6,1,0.4082	0.9428	0.9642	0.9996	0.9984	0.9950	0.9940	0.9974	0.9512	0.9504	0.9606
len	0.2289	0.3518	0.2849	0.2841	3.3915	3.3915	3.7303	2.4496	2.4496	2.5330
300,6,1,0.9	0.9488	0.9820	0.9986	0.9990	0.9962	0.9922	0.9974	0.9556	0.9416	0.9588
len	0.2294	1.8832	1.6949	1.6951	3.3534	3.3534	3.7269	2.4318	2.4318	2.5545
300,6,4,0	0.9468	0.9500	0.9494	0.9974	0.9970	0.9350	0.9958	0.9426	0.9444	0.9444
len	0.2292	0.2306	0.2308	0.1865	2.1448	2.1448	2.3964	3.3597	3.3597	3.3631
300,6,4,0.4082	0.9460	0.9492	0.9498	0.9974	0.9962	0.9366	0.9954	0.9550	0.9540	0.9620
len	0.2293	0.3502	0.3500	0.2837	2.1425	2.1425	2.3992	3.3865	3.3865	3.4647
300,6,4,0.9	0.9536	0.9466	0.9508	0.9974	0.9962	0.9280	0.9956	0.9550	0.8984	0.9488
len	0.2298	1.9734	1.9803	1.7535	2.1208	2.1208	2.3638	3.4309	3.4309	3.7048

Table 4.8. Bootstrapping OLS Forward Selection with C_p, $e_i \sim EXP(1) - 1$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
175,7,1,0	0.9426	0.9528	0.9972	0.9980	0.9956	0.9948	0.9970	0.9376	0.9354	0.9374
len	0.2988	0.3017	0.2445	0.2449	3.6585	3.6585	4.0358	2.4531	2.4531	2.4628
175,7,1,0.3780	0.9492	0.9614	0.9970	0.9980	0.9942	0.9920	0.9978	0.9468	0.9474	0.9546
len	0.2989	0.4465	0.3646	0.3628	3.6610	3.6610	4.0213	2.4513	2.4513	2.5268
175,7,1,0.9	0.9498	0.9588	0.9978	0.9980	0.9952	0.9948	0.9974	0.9552	0.9506	0.9616
len	0.2989	2.4006	2.2848	2.2729	3.6830	3.6830	4.0532	2.4683	2.4683	2.6119
175,7,5,0	0.9432	0.9444	0.9432	0.9988	0.9974	0.9334	0.9970	0.9328	0.9324	0.9328
len	0.2987	0.3011	0.3008	0.2426	2.1328	2.1328	2.3991	3.6039	3.6039	3.6080
175,7,5,0.3780	0.9464	0.9562	0.9536	0.9980	0.9972	0.9388	0.9968	0.9490	0.9488	0.9554
len	0.2997	0.4454	0.4445	0.3599	2.1422	2.1422	2.4011	3.6274	3.6274	3.6960
175,7,5,0.9	0.9480	0.9372	0.9414	0.9972	0.9964	0.9188	0.9956	0.9674	0.9306	0.9712
len	0.2999	2.5035	2.4907	2.3207	2.1209	2.1209	2.3672	3.6911	3.6911	4.0296
400,8,1,0	0.9470	0.9466	0.9982	0.9986	0.9960	0.9954	0.9980	0.9420	0.9408	0.9414
len	0.1988	0.1997	0.1601	0.1616	3.8815	3.8815	4.2847	2.4518	2.4518	2.4575
400,8,1,0.3536	0.9500	0.9558	0.9988	0.9990	0.9982	0.9972	0.9990	0.9510	0.9530	0.9576
len	0.1987	0.2886	0.2325	0.2337	3.8767	3.8767	4.2648	2.4500	2.4500	2.5145
400,8,1,0.9	0.9486	0.9786	0.9980	0.9988	0.9962	0.9938	0.9984	0.9396	0.9136	0.9480
len	0.1992	1.7559	1.5115	1.5105	3.8606	3.8606	4.2689	2.4416	2.4416	2.5540
400,8,6,0	0.9462	0.9554	0.9470	0.9972	0.9964	0.9414	0.9962	0.9454	0.9470	0.9472
len	0.1987	0.1996	0.1997	0.1608	2.1416	2.1416	2.3904	3.7942	3.7942	3.7969
400,8,6,0.3536	0.9506	0.9496	0.9494	0.9978	0.9974	0.9378	0.9976	0.9544	0.9572	0.9626
len	0.1987	0.2868	0.2866	0.2315	2.1460	2.1460	2.4041	3.8150	3.8150	3.8757
400,8,6,0.9	0.9504	0.9516	0.9534	0.9980	0.9970	0.9398	0.9964	0.9392	0.8652	0.9326
len	0.1994	1.8010	1.7988	1.5368	2.1182	2.1182	2.3653	3.8600	3.8600	4.1234

Table 4.9. Bootstrapping OLS Forward Selection with C_p, $e_i \sim EXP(1) - 1$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
250,10,1,0	0.9440	0.9480	0.9992	0.9974	0.9970	0.9970	0.9992	0.9434	0.9438	0.9452
len	0.2511	0.2531	0.2045	0.2047	4.3012	4.3012	4.7407	2.4539	2.4539	2.4646
250,10,1,0.3162	0.9486	0.9550	0.9980	0.9968	0.9960	0.9960	0.9988	0.9456	0.9466	0.9524
len	0.2510	0.3508	0.2837	0.2850	4.2982	4.2982	4.7229	2.4520	2.4520	2.5130
250,10,1,0.9	0.9510	0.9662	0.9980	0.9980	0.9980	0.9952	0.9986	0.9548	0.9478	0.9646
len	0.2511	2.1312	1.9553	1.9525	4.3034	4.3034	4.7428	2.4514	2.4514	2.5978
250,10,8,0.9	0.9420	0.9454	0.9426	0.9978	0.9972	0.9312	0.9968	0.9466	0.8528	0.9374
len	0.2519	2.2039	2.2068	1.9800	2.1253	2.1253	2.3707	4.2581	4.2581	4.6138

Table 4.10. Bootstrapping OLS Forward Selection with C_p, $e_i \sim uniform$(-1,1)

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
100,4,1,0	0.9434	0.9422	0.9970	0.9976	0.9926	0.9794	0.9932	0.9350	0.9366	0.9378
len	0.2287	0.2305	0.1881	0.1869	2.6915	2.6915	2.9950	2.4451	2.4451	2.4521
100,4,1,0.5	0.9476	0.9716	0.9972	0.9982	0.9920	0.9828	0.9954	0.9564	0.9542	0.9608
len	0.2287	0.3806	0.3108	0.3103	2.6869	2.6869	2.9817	2.4530	2.4530	2.5700
100,4,1,0.9	0.9386	0.9836	0.9980	0.9978	0.9924	0.9430	0.9954	0.9466	0.9272	0.9546
len	0.2287	1.6899	1.5346	1.5403	2.5654	2.5654	2.8198	2.4204	2.4204	2.5411
100,4,2,0	0.9446	0.9490	0.9402	0.9976	0.9966	0.9322	0.9962	0.9344	0.9346	0.9348
len	0.2291	0.2306	0.2308	0.1864	2.1338	2.1338	2.3996	2.7908	2.7908	2.7957
100,4,2,0.5	0.9484	0.9552	0.9500	0.9964	0.9952	0.9298	0.9948	0.9562	0.9560	0.9652
len	0.2293	0.3803	0.3801	0.3108	2.1394	2.1394	2.3961	2.8385	2.8385	2.9541
100,4,2,0.9	0.9462	0.9574	0.9612	0.9974	0.9962	0.9340	0.9960	0.9506	0.9288	0.9600
len	0.2298	1.8403	1.8426	1.6604	2.0891	2.0891	2.2997	2.8623	2.8623	3.0610
300,6,1,0	0.9524	0.9520	0.9980	0.9978	0.9938	0.9912	0.9956	0.9458	0.9460	0.9466
len	0.1328	0.1332	0.1073	0.1076	3.3713	3.3713	3.7321	2.4486	2.4486	2.4538
300,6,1,0.4082	0.9474	0.9634	0.9978	0.9978	0.9954	0.9928	0.9976	0.9530	0.9528	0.9606
len	0.1328	0.2034	0.1651	0.1644	3.3757	3.3757	3.7156	2.4459	2.4459	2.5311
300,6,1,0.9	0.9518	0.9668	0.9992	0.9986	0.9932	0.9854	0.9962	0.9470	0.9516	0.9562
len	0.1328	1.2101	0.9844	0.9873	3.3965	3.3965	3.7372	2.4642	2.4642	2.5940
300,6,4,0	0.9484	0.9482	0.9504	0.9988	0.9972	0.9370	0.9980	0.9456	0.9470	0.9468
len	0.1330	0.1331	0.1332	0.1074	2.1397	2.1397	2.3944	3.3324	3.3324	3.3348
300,6,4,0.4082	0.9512	0.9520	0.9478	0.9984	0.9976	0.9378	0.9976	0.9590	0.9600	0.9670
len	0.1330	0.2023	0.2024	0.1641	2.1351	2.1351	2.3891	3.3628	3.3628	3.4424
300,6,4,0.9	0.9492	0.9348	0.9316	0.9962	0.9958	0.9382	0.9952	0.9580	0.9214	0.9476
len	0.1330	1.2383	1.2381	0.9968	2.1263	2.1263	2.3744	3.4591	3.4591	3.5872

Table 4.11. Bootstrapping OLS Forward Selection with C_p, $e_i \sim uniform$(-1,1)

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
175,7,1,0	0.9510	0.9458	0.9976	0.9978	0.9942	0.9936	0.9970	0.9404	0.9410	0.9422
len	0.1737	0.1742	0.1410	0.1410	3.6229	3.6229	4.0099	2.4465	2.4465	2.4560
175,7,1,0.3780	0.9466	0.9594	0.9982	0.9974	0.9922	0.9908	0.9960	0.9496	0.9478	0.9540
len	0.1735	0.2581	0.2101	0.2104	3.6232	3.6232	3.9908	2.4456	2.4456	2.5217
175,7,1,0.9	0.9462	0.9790	0.9984	0.9982	0.9908	0.9852	0.996	0.9372	0.8882	0.9364
len	0.1737	1.5558	1.3132	1.3099	3.6126	3.6126	3.9794	2.4412	2.4412	2.5470
175,7,5,0	0.9468	0.9448	0.9462	0.9984	0.9974	0.9400	0.9974	0.9368	0.9388	0.9388
len	0.1738	0.1746	0.1745	0.1409	2.1341	2.1341	2.3927	3.5503	3.5503	3.5543
175,7,5,0.3780	0.9518	0.9446	0.9410	0.9974	0.9960	0.9282	0.9956	0.9446	0.9452	0.9516
len	0.1736	0.2569	0.2567	0.2079	2.1299	2.1299	2.3853	3.5746	3.5746	3.6444
175,7,5,0.9	0.9510	0.9514	0.9464	0.9972	0.9962	0.9350	0.9956	0.9210	0.8716	0.9186
len	0.1745	1.6019	1.6030	1.3428	2.1150	2.1150	2.3705	3.6267	3.6267	3.8422
400,8,1,0	0.9468	0.9474	0.9982	0.9980	0.9976	0.9980	0.9992	0.9398	0.9396	0.9418
len	0.1151	0.1153	0.0932	0.0931	3.8667	3.8667	4.2677	2.4481	2.4481	2.4538
400,8,1,0.3536	0.9504	0.9624	0.9982	0.9988	0.9962	0.9954	0.9976	0.9508	0.9514	0.9560
len	0.1151	0.1665	0.1347	0.1346	3.8608	3.8608	4.2459	2.4467	2.4467	2.5105
400,8,1,0.9	0.9554	0.9566	0.9988	0.9986	0.9974	0.9950	0.9978	0.9586	0.9654	0.9682
len	0.1151	1.0934	0.8772	0.8733	3.8731	3.8731	4.2550	2.4653	2.4653	2.5422
400,8,6,0	0.9566	0.9486	0.9532	0.9978	0.9976	0.9394	0.9968	0.9442	0.9442	0.9440
len	0.1151	0.1152	0.1153	0.0932	2.1303	2.1303	2.3849	3.7618	3.7618	3.7646
400,8,6,0.3536	0.9434	0.9500	0.9512	0.9988	0.9982	0.9332	0.9980	0.9550	0.9544	0.9614
len	0.1151	0.1654	0.1657	0.1338	2.1423	2.1423	2.3935	3.7822	3.7822	3.8434
400,8,6,0.9	0.9510	0.9334	0.9250	0.9990	0.9976	0.9390	0.9972	0.9646	0.9424	0.9614
len	0.1152	1.1023	1.0997	0.8773	2.1285	2.1285	2.3764	3.8963	3.8963	3.9966

Table 4.12. Bootstrapping OLS Forward Selection with C_p, $e_i \sim uniform$(-1,1)

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
250,10,1,0	0.9498	0.9544	0.9976	0.9982	0.9958	0.9964	0.9996	0.9472	0.9480	0.9498
len	0.1455	0.1459	0.1190	0.1171	4.2680	4.2680	4.7117	2.4484	2.4484	2.4586
250,10,1,0.3162	0.9468	0.9556	0.9974	0.9974	0.9950	0.9952	0.9980	0.9490	0.9472	0.9522
len	0.1452	0.2020	0.1637	0.1647	4.2553	4.2553	4.6901	2.4474	2.4474	2.5071
250,10,1,0.9	0.9520	0.9598	0.9988	0.9986	0.9964	0.9924	0.9980	0.9390	0.9156	0.9384
len	0.1455	1.3920	1.1266	1.1317	4.2650	4.2650	4.6935	2.4598	2.4598	2.5564
250,10,8,0.9	0.9480	0.9344	0.9366	0.9988	0.9970	0.9376	0.9968	0.9170	0.8658	0.9076
len	0.1462	1.4146	1.4169	1.1037	2.1692	2.1692	2.4177	4.2472	4.2472	4.4169

Table 4.13. Bootstrapping OLS Forward Selection with C_p, $e_i \sim 0.9N(0,1) + 0.1N(0,100)$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
100,4,1,0	0.9430	0.9426	0.9988	0.9984	0.9950	0.9816	0.9954	0.9488	0.9354	0.9486
len	1.2753	1.3079	1.0849	1.0807	2.7354	2.7354	3.0313	2.5185	2.5185	2.5453
100,4,1,0.5	0.9372	0.9626	0.9976	0.9978	0.9906	0.9706	0.9930	0.9590	0.9452	0.9632
len	1.2795	1.9701	1.8551	1.8475	2.7040	2.7040	2.9829	2.4952	2.4952	2.6173
100,4,1,0.9	0.9430	0.9258	0.9966	0.9966	0.9870	0.9842	0.9938	0.9698	0.9690	0.9738
len	1.2615	9.0221	9.0443	8.9747	2.7744	2.7744	3.0145	2.5924	2.5924	2.7313
100,4,2,0	0.9358	0.9380	0.9396	0.9974	0.9964	0.9278	0.9960	0.9412	0.9128	0.9362
len	1.2762	1.3133	1.3141	1.0669	2.1429	2.1429	2.4031	2.9546	2.9546	3.0108
100,4,2,0.5	0.9376	0.9518	0.9526	0.9980	0.9972	0.9244	0.9962	0.9636	0.9456	0.9692
len	1.2772	2.0428	2.0525	1.8828	2.1431	2.1431	2.3914	2.9581	2.9581	3.1646
100,4,2,0.9	0.9418	0.9036	0.9030	0.9960	0.9942	0.9556	0.9942	0.9860	0.9850	0.9904
len	1.2705	8.8549	8.8536	8.8718	2.1365	2.1365	2.3406	3.1181	3.1181	3.3162
300,6,1,0	0.9502	0.9502	0.9988	0.9988	0.9956	0.9942	0.9972	0.9538	0.9562	0.9562
len	0.7541	0.7676	0.6125	0.6163	3.4313	3.4313	3.7789	2.4803	2.4803	2.4857
300,6,1,0.4082	0.9536	0.9528	0.9994	0.9990	0.9954	0.9918	0.9978	0.9590	0.9516	0.9610
len	0.7518	1.1835	0.9440	0.9431	3.4470	3.4470	3.7793	2.4964	2.4964	2.5875
300,6,1,0.9	0.9408	0.8938	0.9994	0.9988	0.9968	0.9964	0.9992	0.9734	0.9724	0.9772
len	0.7572	5.7952	5.8075	5.8075	3.4117	3.4117	3.7150	2.5554	2.5554	2.7015
300,6,4,0	0.9426	0.9438	0.9440	0.9986	0.9976	0.9340	0.9972	0.9694	0.9708	0.9714
len	0.7535	0.7655	0.7662	0.6117	2.1399	2.1399	2.3965	3.4677	3.4677	3.4741
300,6,4,0.4082	0.9474	0.9416	0.9408	0.9980	0.9972	0.9376	0.9974	0.9708	0.9490	0.9682
len	0.7540	1.2049	1.2032	0.9567	2.1352	2.1352	2.3887	3.5393	3.5393	3.6342
300,6,4,0.9	0.9470	0.8442	0.8516	0.9990	0.9976	0.9204	0.9972	0.9950	0.9942	0.9976
len	0.7566	5.5209	5.4942	5.4929	2.1338	2.1338	2.4184	3.5959	3.5959	3.9241

Table 4.14. Bootstrapping OLS Forward Selection with C_p, $e_i \sim 0.9N(0,1) + 0.1N(0,100)$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
175,7,1,0	0.9436	0.9334	0.9976	0.9988	0.9960	0.9954	0.9984	0.9502	0.9486	0.9530
len	0.9782	1.0132	0.8002	0.8041	3.7347	3.7347	4.0999	2.5087	2.5087	2.5264
175,7,1,0.3780	0.9448	0.9548	0.9986	0.9988	0.9946	0.9906	0.9964	0.9468	0.9236	0.9472
len	0.9760	1.4558	1.1879	1.1995	3.7414	3.7414	4.0960	2.5019	2.5019	2.5990
175,7,1,0.9	0.9460	0.8386	0.9998	0.9974	0.9974	0.9974	0.9982	0.9754	0.9754	0.9806
len	0.9739	7.6005	7.6388	7.6030	3.6890	3.6890	4.0249	2.5665	2.5665	2.7239
175,7,5,0	0.9438	0.9360	0.9306	0.9982	0.9978	0.9332	0.9974	0.9616	0.9498	0.9576
len	0.9824	1.0211	1.0213	0.8030	2.1507	2.1507	2.4162	3.8365	3.8365	3.8635
175,7,5,0.3780	0.9420	0.9492	0.9444	0.9982	0.9972	0.9352	0.9962	0.9550	0.9156	0.9472
len	0.9838	1.4940	1.4923	1.2229	2.1325	2.1325	2.3886	3.8182	3.8182	3.9628
175,7,5,0.9	0.9426	0.8104	0.8104	0.9976	0.9968	0.9246	0.9966	0.9950	0.9934	0.9974
len	0.9796	7.3776	7.4012	7.3634	2.1353	2.1353	2.4001	3.9222	3.9222	4.2738
400,8,1,0	0.9546	0.9506	0.9982	0.9984	0.9962	0.9966	0.9982	0.9582	0.9600	0.9598
len	0.6547	0.6600	0.5316	0.5269	3.9272	3.9272	4.3193	2.4727	2.4727	2.4778
400,8,1,0.3536	0.9526	0.9484	0.9986	0.9992	0.9972	0.9968	0.9992	0.9632	0.9648	0.9688
len	0.6550	0.9705	0.7660	0.7696	3.9382	3.9382	4.3146	2.4872	2.4872	2.5529
400,8,1,0.9	0.9488	0.8574	0.9986	0.9992	0.9992	0.9988	0.9994	0.9724	0.9716	0.9776
len	0.6539	5.1142	5.1091	5.0954	3.9079	3.9079	4.2759	2.5505	2.5505	2.7119
400,8,6,0	0.9494	0.9516	0.9474	0.9982	0.9978	0.9322	0.9972	0.9612	0.9616	0.9620
len	0.6525	0.6579	0.6585	0.5302	2.1374	2.1374	2.3880	3.8894	3.8894	3.8929
400,8,6,0.3536	0.9528	0.9368	0.9394	0.9986	0.9978	0.9354	0.9980	0.9750	0.9676	0.9752
len	0.6546	0.9710	0.9723	0.7721	2.1371	2.1371	2.3909	3.9903	3.9903	4.0548
400,8,6,0.9	0.9446	0.8708	0.8668	0.9998	0.9988	0.9408	0.9980	0.9940	0.9908	0.9962
len	0.6569	5.1147	5.1245	5.0298	2.1306	2.1306	2.3661	4.0631	4.0631	4.4286

Table 4.15. Bootstrapping OLS Forward Selection with C_p, $e_i \sim 0.9N(0,1) + 0.1N(0,100)$

n,p,k,ψ	β_1	β_2	β_{p-1}	β_p	pm0	hyb0	br0	pm1	hyb1	br1
250,10,1,0	0.9416	0.9372	0.9982	0.9986	0.9988	0.9988	0.9994	0.9484	0.9512	0.9526
len	0.8268	0.8477	0.6728	0.6716	4.3928	4.3928	4.8144	2.4911	2.4911	2.5040
250,10,1,0.3162	0.9478	0.9404	0.9982	0.9984	0.9982	0.9978	0.9990	0.9528	0.9448	0.9538
len	0.8197	1.1727	0.9306	0.9261	4.4017	4.4017	4.8054	2.5024	2.5024	2.5688
250,10,1,0.9	0.9464	0.8054	0.9990	0.9970	0.9982	0.9988	0.9996	0.9688	0.9660	0.9736
len	0.8216	6.4908	6.4570	6.4595	4.3751	4.3751	4.7827	2.5690	2.5690	2.7350
250,10,8,0.9	0.9504	0.8224	0.8338	0.9992	0.9984	0.9406	0.9976	0.9968	0.9962	0.9992
len	0.8243	6.4905	6.4930	6.4301	2.1402	2.1402	2.3876	4.5258	4.5258	4.9284

Suppose $\psi = 0$. Then from Section 3, $\hat{\boldsymbol{\beta}}_S$ has the same limiting distribution for I_{min} and the full model. Note that the average lengths and coverages for forward selection I_{min} CIs for β_1 and β_2 were close to the expected full model lengths $3.92/\sqrt{n}$. The lengths were shorter for β_{p-1} and β_p. For $\psi \geq 0$, the I_{min} coverages were higher than 0.95 for the inactive predictors (and for the tests pm0, hyb0 and br0) since zeros often occurred for inactive $\hat{\beta}_j^*$.

CHAPTER 5

CONCLUSIONS

There is massive literature on variable selection and a fairly large literature for inference after variable selection. See references in Pelawa Watagoda and Olive (2018).

Response plots of the fitted values $\hat{Y}$ versus the response Y are useful for checking linearity of the MLR model and for detecting outliers. Residual plots should also be made.

The simulations were done in *R*. See R Core Team (2016). We used several *R* functions including forward selection as computed with the `regsubsets` function from the `leaps` library. The collection of Olive (2018b) *R* functions *slpack*, available from (http://lagrange.math.siu.edu/Olive/slpack.txt), has some useful functions for the inference. Table 1 was made with `vsbootsim4`. There was occasional undercoverage for the shorth and hybrid region, especially when $\psi = 0.9$.

REFERENCES

Akaike, H. (1973), "Information Theory as an Extension of the Maximum Likelihood Principle," in *Proceedings, 2nd International Symposium on Information Theory*, eds. Petrov, B.N., and Csakim, F., Akademiai Kiado, Budapest, 267-281.

Bickel, P.J., and Ren, J.–J. (2001), "The Bootstrap in Hypothesis Testing," in *State of the Art in Probability and Statistics: Festschrift for William R. van Zwet*, eds. de Gunst, M., Klaassen, C., and van der Vaart, A., The Institute of Mathematical Statistics, Hayward, CA, 91-112.

Büchlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927-961.

Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods* & *Research*, 33, 261-304.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans,* SIAM, Philadelphia, PA.

Efron, B. (2014), "Estimation and Accuracy After Model Selection," (with discussion), *Journal of the American Statistical Association,* 109, 991-1007.

Freedman, D.A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218-1228.

Friedman, J.H., and Hall, P. (2007), "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference,* 137, 669-683.

Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference,* 143, 1039-1048.

Hall, P. (1988), "Theoretical Comparisons of Bootstrap Confidence Intervals," (with discussion), *The Annals of Statistics*, 16, 927-985.

Hebbler, B. (1847), "Statistics of Prussia," *Journal of the Royal Statistical Society, A*, 10, 154-186.

Knight, K., and Fu, W.J. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378.

Machado, J.A.F., and Parente, P. (2005), "Bootstrap Estimation of Covariance Matrices Via the Percentile Method," *Econometrics Journal,* 8, 70–78.

Nishi, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *The Annals of Statistics,* 12, 758-765.

Olive, D.J. (2013), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2018a), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers*, to appear, see (http://lagrange.math.siu.edu/Olive/pphpr.pdf).

Olive, D.J. (2018b), *Prediction and Statistical Learning*, online course notes, see (http://lagrange.math.siu.edu/Olive/slearnbk.htm).

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2018), "Inference for Multiple Linear Regression After Model or Variable Selection," preprint at (http://lagrange.math.siu.edu/Olive/ppvsinf.pdf).

R Core Team (2016), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461-464.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* Wiley, New York, NY.

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate

Linear Regression," *Biometrika*, 99, 687-702.

VITA

Graduate School
Southern Illinois University

Drew C. Imhoff

dcimhoff87@gmail.com

Southern Illinois University
Bachelor of Science, Mathematics, Dec 2016

Research Paper Title:
Bootstrapping Forward Selection With C_p

Major Professor: Dr. David J. Olive